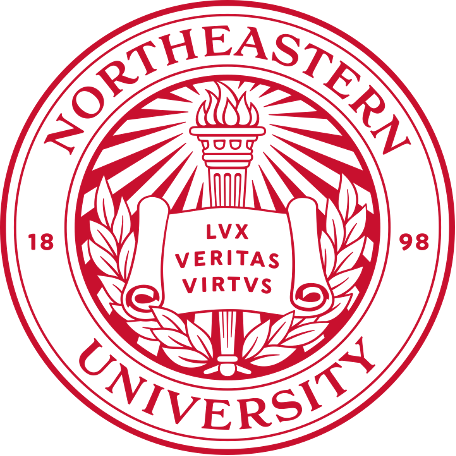
****

**EAI 6010 70915 Applications of Artificial Intelligence**

**Instructor: Mitch Harris**

**Module 5 Assignment**

**Microservice for a model**

**Submission Date: 12/07/2024**

**Jinsheng Wei**

[**wei.jins@northeastern.edu**](mailto:wei.jins@northeastern.edu)

#### Introduction

This document details the implementation and deployment of a microservice for a sentiment analysis classification model, building upon the concepts introduced in Module 5 Lesson 2 (Deploying Models). While the lesson provided a structured approach to model deployment through Google Cloud Platform (GCP), the practical implementation journey led to the discovery of a more streamlined solution using Hugging Face Spaces and Gradio.

The initial implementation attempt followed Module 5 Lesson 2's recommended path, utilizing GCP for deployment - specifically incorporating Cloud Storage for model hosting, Dockerfiles for containerization, and the Google Cloud CLI for service deployment. However, technical challenges with service availability and dependency management, coupled with the rapid evolution of cloud services, necessitated exploring alternative deployment strategies.

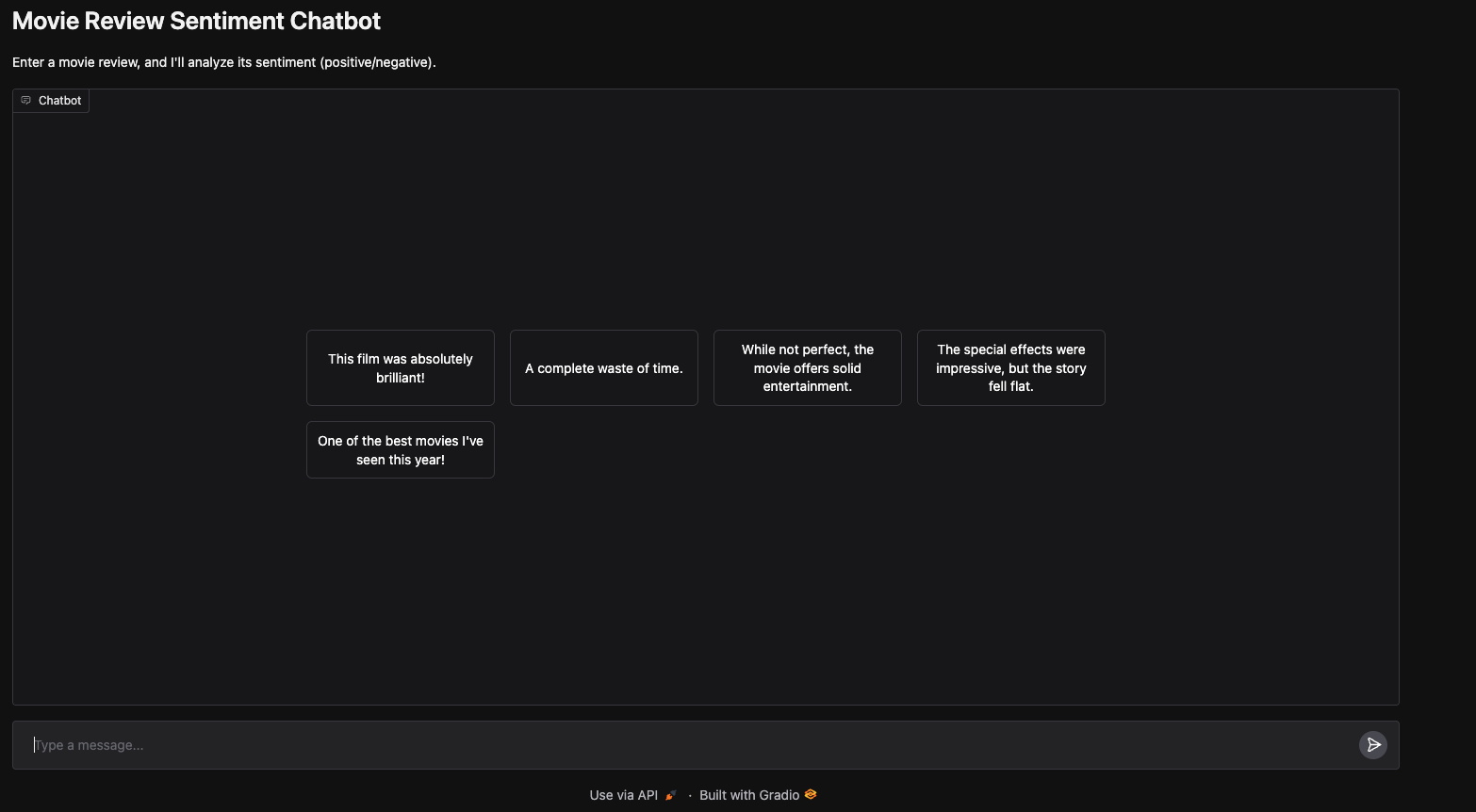
This report presents a successful implementation of a movie review sentiment analysis classifier as a microservice. By leveraging Hugging Face Spaces for deployment and Gradio for creating an interactive interface, the service provides users with a simple yet effective way to input movie reviews and receive sentiment classification results. This approach not only ensures reliable service availability but also offers a user-friendly interface that makes the model accessible to end users.

The following sections detail the service's input/output specifications, provide specific usage examples, and include the necessary URL information for accessing the deployed model.

#### Model Description

The sentiment analysis service is built upon the DistilBERT model, which is a distilled version of BERT optimized for faster inference while maintaining strong performance. Our specific implementation uses a fine-tuned version of distilbert/distilbert-base-uncased, which underwent two iterations of training. The model was initially trained on the IMDB dataset for binary sentiment classification (positive or negative) and further fine-tuned on the "Rotten Tomatoes" dataset to enhance its performance on movie-related content.

This fine-tuning process yielded impressive results, with the model achieving an accuracy of 83.96%, precision of 82.67%, and recall of 85.93%, resulting in an F1 score of 84.27%. The model's architecture is specifically optimized for sentiment analysis in movie reviews, making it particularly effective for analyzing movie-related textual content while maintaining efficient inference times.

****

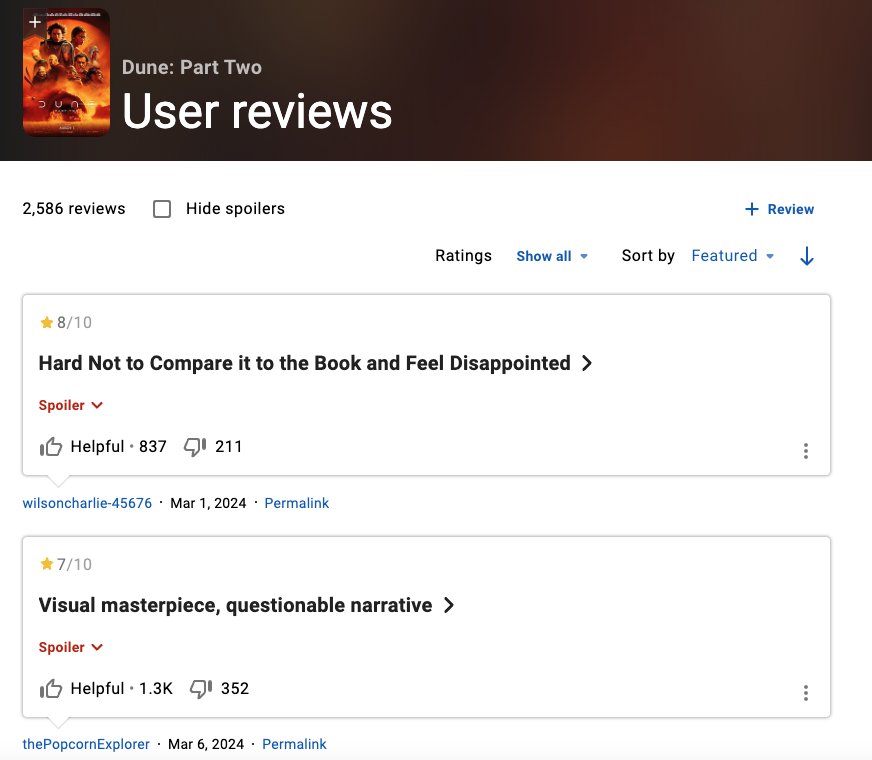
#### Analysis

1. Descriptions of the service’s general input and output.

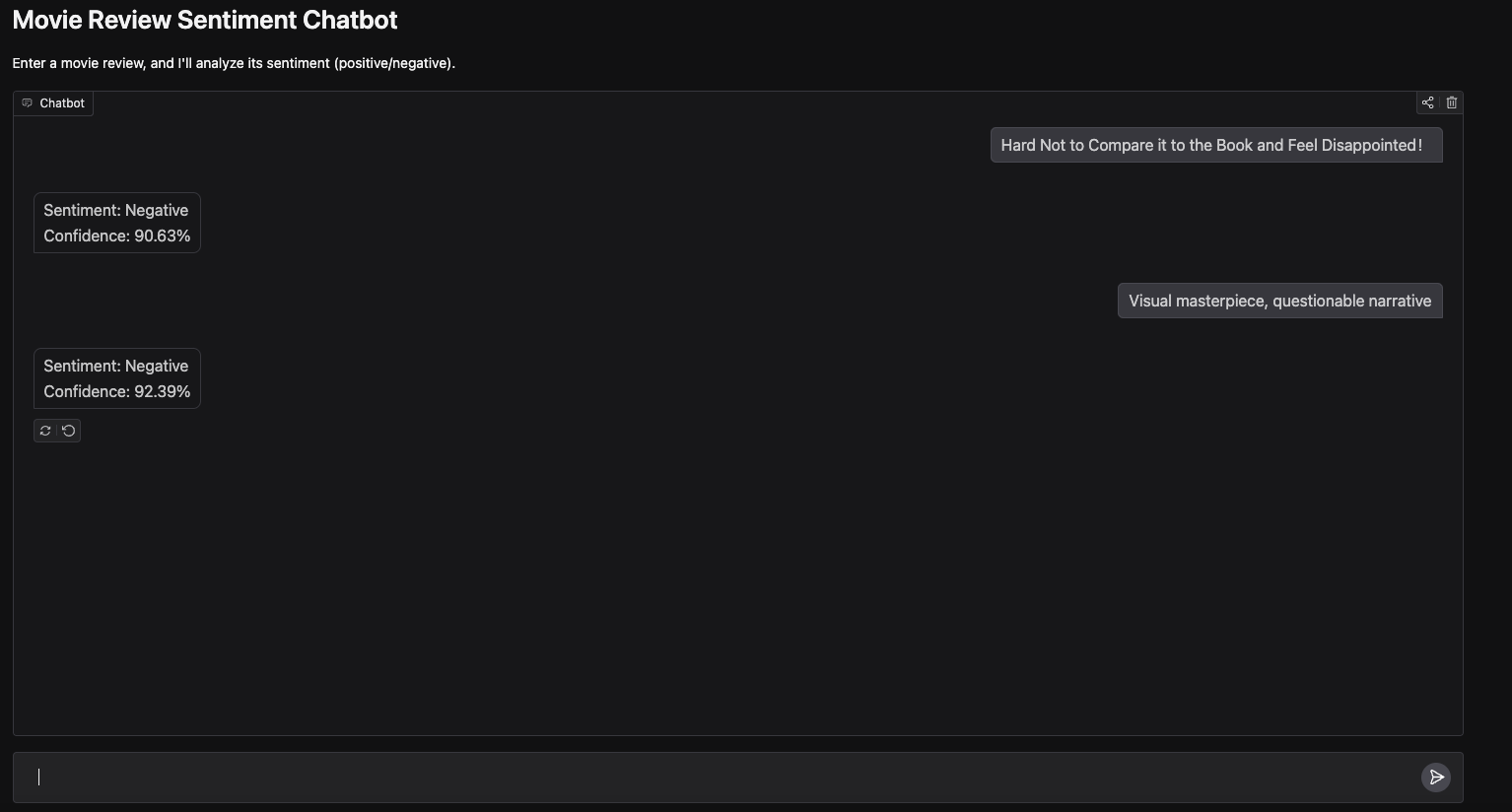
The service is designed to accept movie reviews or movie-related comments as text input. Users can input any length of English text containing their movie review or comment, and the text should be plain without any special formatting or HTML tags.

The service outputs a binary sentiment classification result, simply indicating whether the input review is "Positive" or "Negative". This straightforward output format makes it easy for users to quickly understand the predicted sentiment of their movie review.

1. Specific examples of the service’s input and output.



Let's demonstrate our sentiment analysis service using two recent reviews of "Dune: Part Two" from IMDB:



For the first example, when input with "Hard Not to Compare it to the Book and Feel Disappointed", our model classifies it as a negative sentiment with a high confidence of 90.63%.

Similarly, for the second review "Visual masterpiece, questionable narrative", the model also predicts a negative sentiment with an even higher confidence of 92.39%. These examples showcase how our model accurately captures the critical undertones in movie reviews, even when they contain mixed sentiments about different aspects of the film.

1. The URL of the service.

The sentiment analysis service is deployed as an interactive chatbot interface via Hugging Face Spaces and can be accessed at<https://huggingface.co/spaces/Jasonwei4/week5_assignment_jinshengw>. Built using Gradio, it provides a user-friendly interface where users can directly input their movie reviews. The interface includes example prompts like "This film was absolutely brilliant!" and "The special effects were impressive, but the story fell flat" to guide users on how to interact with the service.

Feel free to test out the service yourself and provide any feedback for future improvements!

#### Reference:

*Tashrifmahmud/sentiment\_analysis\_model\_v2 · hugging face*. (2024, January 4). <https://huggingface.co/tashrifmahmud/sentiment_analysis_model_v2>

*Stanfordnlp/imdb · datasets at hugging face*. (2024, August 28). <https://huggingface.co/datasets/stanfordnlp/imdb>

*Cornell-movie-review-data/rotten\_tomatoes · datasets at hugging face*. (2024, August 28). <https://huggingface.co/datasets/cornell-movie-review-data/rotten_tomatoes>

*Dune: Part two (2024) - user reviews - imdb*. (n.d.). [Video recording]. Retrieved 7 December 2024, from https://www.imdb.com/title/tt15239678/reviews/